Impact of human genome sequencing for *in silico* target discovery

Philippe Sanseau

The year 2000 stands as a landmark in modern biology: the first draft of the human genome sequence has been completed. For the pharmaceutical industry, this achievement provides tremendous opportunities because the genomic sequence exposes all human drug targets for therapeutic intervention. The challenge for the pharmaceutical companies is to exploit this definitive resource for the identification of potential molecular targets, rapid characterization of their function and validation of their involvement in disease pathology. Bioinformatics approaches provide increasingly crucial tools to systematically support this exploratory target drug discovery activity.

Philippe Sanseau
Target Bioinformatics,
GlaxoSmithKline
Gunnels Wood Road
Stevenage
UK SG1 2NY
tel: +44 1438 768119
fax: +44 1438 768097
e-mail: ps14446@
glaxowellcome.co.uk

▼ The recent announcement of the completion of the first draft of the human genome sequence, by a consortium of 16 public laboratories, is clearly a major achievement in biology. Although there are ongoing efforts to produce a high-quality finished sequence by 2003, the draft data can already be used as a resource to identify the majority of the human genes. For the pharmaceutical industry, this is a ‘once in a lifetime’ opportunity to identify all human targets. The challenge is not only to identify the genes and establish long lists of transcripts, but to pick the winners for drug discovery efforts. To support this process, bioinformatics approaches are playing an increasingly crucial role. In addition to aiding the gene identification process, they also help in the characterization phase by using *in silico* analysis.

Nature of the human genomic sequence

In 1995, an international public consortium outlined the plan to complete the human genome sequence. The agreed strategy was to use genomic clones from a physical map and to subject the chosen clones to shotgun sequencing. Several rounds of sequencing are necessary to achieve the target of 99.99% accuracy and ten times coverage for the high quality or finished sequence. Because of the progress in sequencing technology and in response to the pressure from private sequencing initiatives, the public consortium decided to accelerate the release of a working draft sequence. In comparison to the finished reference sequence, the draft has a depth of 3–4 times coverage and contains sequence contigs as small as 1 kb. The draft data are available to the scientific community within 24 h of assembly. Approximately 20% of the sequence available at the time of writing is of finished quality. Although the majority of data exist as draft sequence, this covers approximately 95% of the genome, and is therefore an important resource for gene identification and characterization. The nature of the human genome sequence has advantages but also drawbacks (Box 1). Because the draft data are ‘dynamic’, tracking of the sequence fragments is a challenge. Public initiatives, such as EnsEMBL (http://www.ensembl.org) at the European Bioinformatics Institute (EBI; Cambridge, UK), attempt to provide the tracking framework and gene annotations to solve such issues.

Since December 2000, the sequences of the two smallest human chromosomes have been published in the public domain: chromosomes 22 (Ref. 1) and 21 (Ref. 2). In both cases, more than 33 Mb of DNA have been sequenced as finished sequence.

In contrast to the public effort, the private company Celera (Rockville, MD, USA) uses a whole-genome shotgun sequencing[3,4] strategy. In this case, a single random sheared genomic DNA library is prepared and fragments are cloned into vectors such as plasmids. A large number of insert ends are sequenced to ensure complete coverage of the genome and the sequences are reassembled using computer

1359-6446/01/$ – see front matter ©2001 Elsevier Science Ltd. All rights reserved. PII: S1359-6446(01)01682-8

Box 1. Advantages and disadvantages of using genomic sequence for computational gene identification

Advantages
Sequence quality
All transcripts (including rare genes) will be present
Full gene structure might be available
Access to regulatory regions of the genes

Disadvantages
Dynamic and fragmented data
Need to clone the gene
Contamination
No expression information
Difficult to identify splice variants and regulatory regions

programs. This strategy has been used successfully in the past for microbial genomes such as *Haemophilus influenzae*[5] and, more recently, for larger genomes such as *Drosophila melanogaster*[6]. At the same time as the public consortium, in June 2000, Celera announced the first assembly of the human genome sequence using the whole-genome shotgun sequencing strategy. Private companies can supplement their proprietary data with public domain maps and sequence information.

How many opportunities?

The human genome sequence contains sufficient information to identify novel opportunities or genes for therapeutic intervention. Until recently, the main source of novelty came from expressed sequence tags (ESTs)[7,8]. EST collections have proven to be useful for identifying novel targets, for example, cathepsin K[9]. However, rare transcripts or genes with a limited pattern of expression, although attractive as drug targets, are the most difficult to identify using EST technology. Because of these limitations it is estimated that 10%–20% of genes are not present in EST databases[8]. The challenge for computational biology is to identify these transcripts from the significant expanses of non-coding sequences from the human genome sequence. One of the most interesting questions for the scientific community is to estimate the total number of human genes.

What is a gene?

Because the definition of a gene varies, so does the total number of estimated human genes. Classically, genes are defined as inheritable units responsible for an observable phenotype. However, the same transcript unit can produce different splice products with different functions, or post-translational modification of proteins. In 1994, Fields and colleagues[10] proposed that genes are 'distinct transcription units or parts of transcription units that can be translated to generate one or a set of related amino acid sequences'. During 10–14 May 2000, at the meeting *Genome Sequencing and Biology* (Cold Spring Harbor, NY, USA), a debate took place on how many genes were present in the human genome. This led to the 'human gene sweep' (http://www.ensembl.org), in which individuals bet on the total number of human genes. This discussion led to yet to another definition of the term gene. For the gene sweep, a gene is defined as a set of connected transcripts, wherein a transcript is a set of exons generated via transcription, possibly followed by pre-mRNA splicing. Transcripts are connected if they share at least part of one exon. If the proteins produced are different as a result of alternative splice events, they will still belong to the same gene.

Total number of human genes

Even with the draft sequence available, assessing the total number of genes in the human genome remains a challenging task. Several groups have recently published estimates ranging from 28,000 to more than 120,000 genes[11–13]. The estimations established with the publication of the sequence of the first two human chromosomes are also well below the 100,000 mark. Both chromosomes represent approximately 2% of the total human genome; if they are 'representative' of other chromosomes, the authors estimate the total number of human genes to be ~40,000 (Refs 1,2). In the year 2000, the full genome sequence of the laboratory fruit fly (*D. melanogaster*) was published[14]. The total number of predicted genes for *Drosophila* is ~13,600. It was a surprise to see that this number was significantly less than the 19,000 genes identified for an apparently simpler organism, *Caenorhabditis elegans*, which was fully sequenced in 1998 (Ref. 15). This emphasizes the difficulty of accurately predicting the number of transcripts for a given organism. The lower estimate of human genes might reflect the importance of process events, for example, splicing of RNA or protein modifications such as glycosylation, in generating the complexity associated with higher organisms.

Since 10 October 2000, the highest bet in the human gene sweep competition was 200,000 genes and the lowest 27,462, with a mean value of 62,598. The uncertainty surrounding the number of human genes raises the important question: how can a gene be recognized?

Recognition of genes

The first goals that must be achieved before the preliminary prioritization of genes can occur are the computational gene identification and annotation of the human

genome sequence. The programs for gene identification can be split into two broad categories:
(1) Blast[16], Procrustes[17] and Genewise[18], which are based on comparisons of novel genomic sequence to homologous protein sequences; and
(2) the gene prediction software for *ab initio* prediction of genes from genomic sequences such as GenScan[19] and Grail[20].

Recently, Guigo and coworkers[21] published an assessment of several methods. The report concluded that although the performance of the similarity-based programs was not affected by longer genomic sequences, the accuracy of *ab initio* methods such as Genscan dropped significantly. However, the sensitivity of *ab initio* methods remained high. It is also known that these methods can generate up to 30% over-predictions. For the chromosome 22 data, Genscan detected, at least partially, 94% of the annotated genes. Only 20% of the gene predictions had the correct gene structure and 16% of exons in known genes were not detected at all[1]. In summary, *ab initio* techniques can be used to detect a gene but not to accurately annotate the gene structure. Moreover, these programs failed completely in finding non-protein coding genes such as Xist[22]. The recommendation for computational gene identification is to use a range of prediction programs in parallel with similarity-based approaches.

The identification of gene promoters or regulatory regions from genomic sequences is even more challenging[23–26]. An improvement in these methods will be necessary to fully exploit the amount of sequence information available. Knowledge of transcriptional-control sites is of importance to the pharmaceutical industry to develop molecules to regulate gene expression.

What are the best drug targets?

Only a small fraction of the human genes from the human genome will be amenable to therapeutic intervention. Reports published in 1997 estimate that current therapies are based on ~500 molecular targets[27,28]. The majority of these targets are receptors such as G-protein-coupled receptors (GPCRs), with 45% of all targets; enzymes account for 28% of the remainder. Some target classes are, therefore, more ‘successful’ or have a better tractability in the drug discovery process. Therapeutic proteins or protein drugs such as recombinant proteins and monoclonal antibodies must also be considered when identifying novel targets in the human genome. The total number of tractable targets remains difficult to establish given the uncertainty surrounding the total number of human genes. However, it has been estimated that the number of drug targets is probably 5,000–10,000 (Ref. 29).

Other mammalian genomes?

On 6 October 2000, a public announcement declared that the National Institutes of Health (Bethesda, MD, USA), the Wellcome Trust (London, UK) and three private companies had formed the Mouse Sequencing Consortium (MSC), in an effort to accelerate completion of the mouse genome. By March 2001, a draft sequence should be available to the scientific community.

Because humans and rodents share many basic biological functions, it is probable that protein-coding genes of similar functions will share a high degree of sequence identity, not only in exons, but also in regulatory regions. Thus, the availability of the mouse genome sequence will be highly instrumental in interpreting the human genome. Moreover, the mouse genes can be used for functional analysis.

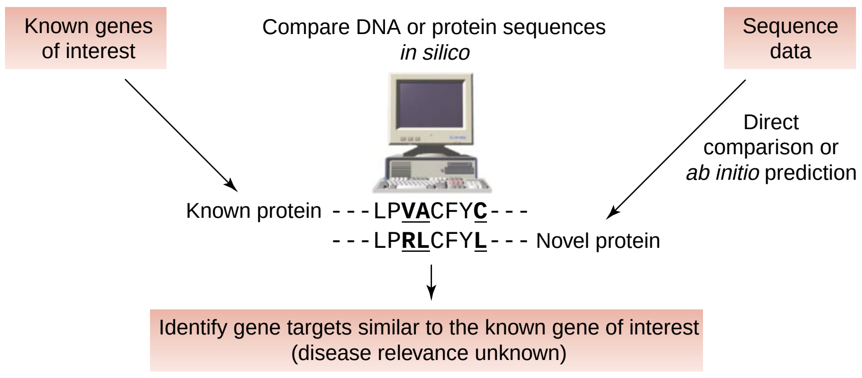
Exploiting the human genome

Discovery genomics: a strategy to identify in silico *novel targets*

Perhaps the simplest strategy to identify potential drug targets from the human genome is to use bioinformatics tools and the numerous sequence databases available, in a process known as ‘discovery genomics’[30] (Fig. 1). Gene predictions from *ab initio* or sequence similarity programs can be used against databases of known effective drug target classes, such as GPCRs, proteases, ion channels, hormone nuclear receptors, kinases, and so on, to uncover novel members of these gene families. Choosing a particular target class will depend on several criteria:

- Has a validated drug target been successfully ‘derived’ from the gene family under study?
- How much experience is available in-house for a particular target class?
- How easy is it to identify ligands for receptors or substrates for enzymes?
- How much downstream work will be necessary to obtain a full-length sequence? (For example, it is well known that some GPCRs are relatively small genes with a single exon, whereas the majority of ion channels have a much more complex gene structure.)

Depending on the nature of the target class, conserved motifs or full-length sequences can be used for the searches. Using this approach, it is relatively easy to identify novel genes belonging to high-value, therapeutic target classes. For example, a novel large family of human taste-receptors belonging to the GPCR gene class was recently identified[31] by computational analysis of sequences obtained from genomic clones. SmithKline Beecham (Harlow, UK) identified more than 100 novel human GPCR orphan receptors[32] using a discovery genomics approach by screening databases.


Known genes of interest
Compare DNA or protein sequences *in silico*
Sequence data
Direct comparison or *ab initio* prediction
Known protein - - - LP**VA**CFY**C** - - -
- - - LP**RL**CFY**L** - - - Novel protein
Identify gene targets similar to the known gene of interest (disease relevance unknown)

Drug DiscoveryToday

Figure 1. The discovery genomics process. In this approach, sequences of known genes of interest such as G-protein-coupled receptors (GPCRs) can be either computationally directly compared with the genome sequence data or against *ab initio* predicted genes from the genome sequence. In the first case, programs such as Blast or GeneWise can be used. For *ab initio* prediction, programs such as Genscan or Grail are applied in a preliminary step to identify open reading frames sequences from the genomic sequence. In a second step, these sequences are compared with known genes of interest by similarity programs. In both cases, the novel genes identified are not directly associated with a disease. They might not be full length and are only known to be similar to the genes of interest used as baits. In the example described, amino acid sequences are compared.

In silico methodology can also aid target validation of orphan genes by highlighting the ligand structures that are most likely to associate with an orphan receptor. For example, using standard similarity search tools played a major role in the identification of the human orthologue of rat preprocortistatin[33]. Moreover, computational methods such as phylogenetic analysis can predict ligand–receptor pairing by tree comparison[34] in addition to classification of genes within a target class. Discovery genomics using computational tools will also have a major impact on other classes of drug targets, for example, proteases. Recent estimates predict a total number of 700–1100 human proteases[35] compared with less than 400 mRNAs represented in GenBank. Using bioinformatics tools to compare known and novel proteases, it is possible to assign a novel sequence to a class of proteases. The challenge for both target validation and for understanding the biochemical function of novel proteases will be the identification of the correct substrate(s) and inhibitors.

If novel genes are identified as fragments (i.e. not full length), *in silico* approaches based on sequence similarities can be used to identify the full-length gene. Additionally, if the partial gene under study is represented in the EST databases, it is possible to assemble the full-length gene computationally by applying various clustering techniques[36]. Such an approach was illustrated when the full-length sequence of a novel protein kinase was obtained exclusively *in silico* by clustering ESTs[37] before being verified by classical methods. If only a few ESTs are available, computational clustering, in conjunction with bench work, is the option of choice. Such an approach was used to obtain the complete sequence of the human vanilloid receptor VR1[38], a potential drug target for pain.

Clearly, novel genes are not related to a disease at the identification stage. For example, if they are derived from genomic sequence information only (with no EST data), tissue distribution will be unknown and cannot be used as a clue to prioritize novel genes belonging to a target class. In such instances, novel genes can be mapped (using publicly available mapping resources) and this information in turn is used to identify 'neighboring' disease loci via databases such as the Online Mendelian Inheritance in Man (OMIM)[39] (http://www.ncbi.nlm.nih.gov/Omim). Depending on the gene under investigation, it can be possible to hypothesize a link (and/or mechanism) with a known disease. Verification of this hypothesis will require extensive screening and access to a human population with the relative phenotypes.

In addition, *in silico* methods based on sequence similarity have been developed to identify protein function. A general function was assigned to more than half of 2557 previously uncharacterized yeast proteins[40] using properties such as similar phylogenetic profiles, correlated mRNA expression patterns and patterns of domain fusion. A similar strategy has the potential to be applied to other fully sequenced genomes.

Discovery genetics: an alternative strategy

Discovery genomics using databases and informatics tools is not the only strategy to identify genes of interest to the industry. By using human disease populations and identifying gene variants, it is possible to identify disease susceptibility genes, as shown by the association of one particular allele of the apolipoprotein-E (ApoE) gene with Alzheimer's disease[41]. This process is known as 'discovery genetics'[30]. Using this approach, novel genes will be disease validated but might not belong to a tractable class of targets, as was the case in the discovery genomics strategy. Both strategies make use of similar functional genomics technologies to elucidate the biological role of novel gene sequences. For discovery genetics, functional analysis will focus on the disease allele and might require pathway expansion to identify a tractable gene target.

With the completion of the human genome sequence and annotation initiatives such as EnsEMBL, identification of potential candidate genes in a disease locus is more likely to happen on a computer rather than at the bench. In addition, it has been demonstrated that disease susceptibility genes can be located using SNPs[30,42]. SNPs are single-base differences in a DNA sequence observed between individuals. SNPs can be identified using bench techniques, but in some cases an available alternative is to use *in silico* methodology. For example, SNPs can be uncovered computationally in a cluster of ESTs derived from different cDNA libraries (i.e. potentially from different individuals). Such an approach was used to identify potential SNPs in the ApoE locus[42], although experimental validation was required. Because SNPs are becoming the cornerstone in discovery genetics, a pre-competitive SNP Consortium (TSC) was established in 1999 between pharmaceutical and bioinformatics companies, five academic centres and the Wellcome Trust. The initial goal of the TSC was to identify and map 300,000 SNPs on the human genome. In fact, at least twice this number of SNPs will be available during the first half of 2001. Most of this data can be accessed via a computer and will be integrated with the human genome sequence.

The utility of SNPs will not be restricted to disease gene identification. The availability of SNP maps will make it possible to generate SNP profiles for patients. Such profiles can then be used to select for groups of patients showing adverse responses to a particular drug. The importance of pharmacogenetics and the use of SNP profiles have been reviewed recently[30]. It has also been shown that SNPs in genes can explain adverse responses to drugs. For example, with albuterol, a β-adrenoceptor drug used to treat bronchial asthma, an individual response seems to be associated with a particular SNP in the β_2-adrenoceptor gene[43].

Beyond the human genome sequence: functional genomics and bioinformatics

The challenge does not lie in finding and creating long lists of genes but in validating novel transcripts and picking winners for drug discovery early in the validation process. Various functional genomics platforms and approaches can be used to functionally characterize genes or reveal pathways. Most of them will require bioinformatics support for data management and analysis.

Microarrays

Tissue distribution is one of the key pieces of information in the validation of a target. This can be done by using *in situ* or northern hybridization, reverse transcription-PCR or real-time quantitative PCR (Taqman technology; Genetech, San Francisco, CA, USA and Perkin-Elmer, Foster City, CA, USA)[44,45]. However, high-throughput large-scale genomics approaches are the most popular, for example, oligonucleotide or gene fragment arrays[46–49], wherein the expression of thousands of genes can be monitored simultaneously. The completion of the human genome sequence will offer the possibility to develop arrays containing the full complement of human genes. In the future, splice or polymorphic variants of genes will also be available on arrays; this will enable even greater insight to the mechanics of the human genome.

Microarrays are not limited to gene characterization but can also be applied to toxicogenomics studies[50] or polymorphism detection[51–53]. Because microarrays produce multiple data points for thousands of genes, this platform requires informatics solutions for data management[54], for the tracking of genes and for automatic primer design for oligo-based arrays. In addition, bioinformatics plays an important role in the analysis of the data produced by microarrays. Statistical techniques are being developed to cluster hundreds of genes based on their expression, in an effort to elucidate biological networks. These methods have been used for: (1) the identification of gene expression patterns in human cancer[55–57], (2) responses to chemotherapeutic agents[58] and (3) clustering of genes in the fibroblast response to serum[59]. *In silico* analysis of expression profiles generated by microarrays have also been used to propose putative function to genes[60,61]. In this approach, after computational clustering, unknown genes are assigned a function according to known genes belonging to the same cluster.

With the completion of the human genome sequence, whole-genome expression profiling of all human genes is a genuine possibility. This is already the case for yeast, in which microarrays have been used to exploit genome sequence information. A recent study in yeast has shown that computational analysis of data generated by microarrays can be used not only to assign function to uncharacterized genes but also to identify a target of a drug[62]. The authors constructed a database of yeast reference expression profiles for different mutations or chemical treatments affecting known pathways. These profiles, derived from mutations, were then compared with others caused by a mutation in an uncharacterized gene. This approach assigned eight previously uncharacterized genes to four different cellular pathways. Interestingly, profiles derived from drug-tested cells were similar to profiles generated when drug target genes were mutated. For example, a gene involved in the ergosterol pathway was identified as the previously unknown drug target of the topical anaesthetic dyclonine. Potential functions were verified experimentally. In principle, such strategies could be applied to more complex organisms and will be of great benefit to pharmaceutical research.

Until recently, the use of microarrays has not led to the identification of disease genes. However, two adjacent ABC transporters have been identified as the disease genes in sitosterolaemia[63], a condition characterized by the accumulation of dietary cholesterol. Changes in the level of expression were detected by microarray data analysis and confirmed by the identification of mutations. In addition, the authors used *in silico* approaches such as GenScan analysis on genomic clones to complete the sequence of one of the transporters.

In yeast, microarrays have been used to analyse upstream sequences of co-regulated genes to identify common features[64,65]. Such a strategy applied to the human genome could greatly improve the identification of regulatory regions of human genes.

Proteomics

Characterization of the genome protein complement or proteome (proteomics)[66,67] is another technology that can be used to select candidate gene targets by assigning function to uncharacterized genes. Proteomics not only includes protein identification and characterization by mass spectroscopy[68,69] and two-dimensional gel electrophoresis, but protein–protein interactions identified by technologies such as the yeast two-hybrid system[70]. Proteome analysis is useful because mRNA analysis by microarrays might not reflect the biological function of proteins that perform the tasks in a cell. More recently, protein microarrays have also been developed[71].

Similar to mRNA microarrays, proteomics bioinformatics has an important role to play in data management and data analysis. For example, a large-scale study of protein–protein interactions has been undertaken for the yeast genome[72]. Using a novel bioinformatics platform, the authors visualized several interactions and assigned a biological context for several functionally unclassified proteins. Moreover, the bioinformatics software can be used to enter new sets of data or to identify conserved interactions in other species. As with similar applications, it would be relatively simple to apply this method in the analysis of other species.

In silico methodologies are also being developed to identify protein interactions from genome sequences. For example, 6809 putative protein–protein interactions have been identified in *Escherichia coli* and more than 45,000 have been identified in yeast, and a large number of these interactions are functionally related[73].

Structural genomics

Three-dimensional (3D) protein structures have traditionally been characterized by low-throughput techniques such as X-ray crystallography. Recently, systems to obtain X-ray data in a high-throughput manner have been developed[74]. For structural biologists, it will be a similar change of scale to that experienced by molecular biologists with high-throughput sequencing and microarrays. The high throughput determination of protein structures is a complement to the genome sequence information, and will ultimately lead to a large catalogue of 3D shapes for human proteins. A steady increase in the number of structures available in databases is likely to occur during the next few years.

Progress has also been achieved in the type of proteins amenable to structure resolution. Integral membrane proteins such as receptors and ion channels are good drug targets, however, structure determination for these proteins has been a challenge, until recently, because of technical difficulties such as the production of large amounts of protein. The publication of the structures for a GPCR (Rhodopsin)[75] and a potassium ion channel[76] have demonstrated that such membrane proteins are now amenable to structure determination.

Structural data can be used by drug designers for rational design and for the elimination of the less promising targets. Structural information can also provide insights for the elucidation of protein function because sequence conservation is typically not as strong as structural conservation. 3D structures can provide a putative function to novel genes and also to identify binding sites or catalytic centres. The structure of a protein, AdipoQ, identified by high-throughput sequencing was determined and shown to be similar to TNF-family cytokines[77]. This led to the prediction that the protein might act as a cell-signalling factor. Structure-based genomics could also be used to ascertain function to disease-linked genes identified by positional cloning or by discovery genetics. Tubby-like proteins are a multigene protein family involved in disease phenotypes, such as obesity, retinal degeneration and hearing loss, but no biochemical functionality was assigned to the proteins. The molecular architecture of these proteins, determined by structural analysis, suggested a function in transcriptional modulation[78]; experimental data support this hypothesis. The effects of mutations or polymorphisms on disease susceptibility genes can also be analyzed by structural analysis. For example, structural data have shown that the majority of cancer-causing mutations in the oncogene p53 map to a region in the protein involved in DNA binding[79].

Structural-based genomics is developing rapidly and the establishment of a collaborative industrial consortium, similar to the TSC, has been proposed[80]. The aim of the Structural Genomics Consortium (SGC) will be to develop a public database of structures for key target classes such as enzymes. The goal is to have an operating SGC by early 2001 and to have the funding for three years.

One of the most difficult, but also most important, challenges for bioinformatics in the next few years will be the integration of sequence and functional data derived from computational and experimental approaches. Data tends to come in 'waves' and since the mid 1990s, most of the data have been sequence data. The advantage of genome sequencing is the knowledge that genomes have a limited size. However, this is not the case with high-throughput functional genomics platforms. The number of experiments and therefore data points to be generated has no obvious limits. Moreover, data integration is not an easy task given the lack of current standards and the fact that a large part of the biological data is not internal to companies. In fact, the limited view is to look at biological data only. Intelligent integration with other data domains, such as chemical and clinical, is the ultimate goal. The use of the human genome sequence and gene data, in combination with SNP analysis in response to drug treatments, is an example of the integration of information derived from different domains. This is the step to take to achieve a proper knowledge-discovery environment for pharmaceutical companies.

Conclusion

With the first draft of the human genome available, the identification of the majority of human genes is becoming a genuine possibility. Some of these genes will be similar to known gene drug targets; others will be completely novel. However, in most cases, functional analysis will be a necessity. Bioinformatics together with *in silico* analysis will play a major role in analysing, managing and connecting data to improve functional annotation using genomics methods. At present, computational approaches on their own cannot validate a large number of targets. However, continuous interactions between bench and *in silico* methods are already part of the day-to-day functional analysis of genes.

Acknowledgements

I thank colleagues from the Genome Informatics and Analysis Group at GlaxoSmithKline (Stevenage, UK) for their comments and suggestions.

References

1 Dunham, I. *et al.* (1999) The DNA sequence of human chromosome 22. *Nature* 402, 489–495

2 Hattori, M. *et al.* (2000) The DNA sequence of human chromosome 21. *Nature* 405, 311–319

3 Weber, J.L. and Myers, E.W. (1997) Human whole-genome shotgun sequencing. *Genome Res.* 7, 401–409

4 Venter, J.C. *et al.* (1998) Shotgun sequencing of the human genome. *Science* 280, 1540–1542

5 Fleischmann, R.D. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae. Science* 269, 496–512

6 Myers, E.W. *et al.* (2000) A whole-genome assembly of *Drosophila. Science* 287, 2196–2204

7 Adams, M.D. *et al.* (1993) Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nat. Genet.* 4, 373–380

8 Williamson, A.R. (1999) The Merck Gene Index project. *Drug Discov. Today* 4, 115–122

9 Drake, F.H. *et al.* (1996) Cathepsin K, but not cathepsins B, L or S, is abundantly expressed in human osteoclasts. *J. Biol. Chem.* 271, 12511–12516

10 Fields, C. *et al.* (1994) How many genes in the human genome? *Nat. Genet.* 7, 345–346

11 Ewing, B. and Green, P. (2000) Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.* 25, 232–234

12 Liang, F. *et al.* (2000) Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat. Genet.* 25, 239–240

13 Roest-Crollius, H. *et al.* (2000) Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nat. Genet.* 25, 235–238

14 Adams, M.D. *et al.* (2000) The genome sequence of *Drosophila melanogaster. Science* 287, 2185–2195

15 The *C. elegans* sequencing consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282, 2012–2018

16 Altschul, S.F. *et al.* (1990) Basic alignment search tool. *J. Mol. Biol.* 215, 403–410

17 Gelfand, M.S. *et al.* (1996) Gene recognition via spliced alignment. *Proc. Natl. Acad. Sci. U. S. A.* 93, 9061–9066

18 Birney, E. and Durbin, R. (1997) Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. *ISMB* 5, 56–64

19 Burge, C.B. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78–94

20 Uberbacher, E.C. and Mural, R.J. (1991) Locating protein-coding regions in DNA sequences by a multiple sensor-neural approach. *Proc. Natl. Acad. Sci. U. S. A.* 88, 11261–11265

21 Guigo, R. *et al.* (2000) An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.* 10, 1631–1642

22 Claverie, J.M. (2000) From bioinformatics to computational biology. *Genome Res.* 10, 1277–1279

23 Claverie, J.M. (1998) Computational methods for exon detection. *Mol. Biotechnol.* 10, 27–47

24 Fickett, J.M. and Hatzigeorgiou, A.G. (1997) Eukaryotic promoter recognition. *Genome Res.* 7, 861–878

25 Stormo, G.D. (2000) Gene finding approaches for eukaryotes. *Genome Res.* 10, 394–397

26 Bajic, V. (2000) Comparing the success of different prediction software in sequence analysis: a review. *Briefings in Bioinformatics* 1, 214–228

27 Drews, J. (1997) Genomic sciences and the medicine of tomorrow. In *Human Disease – From Genetic Causes to Biochemical Effects.* (Drews, J. and Ryser, S., eds) pp. 5–9, Blackwell

28 Drews, J. and Ryser, S. (1997) The role of innovation in drug development. *Nat. Biotechnol.* 15, 1318–1319

29 Drews, J. (2000) Drug discovery: a historical perspective. *Science* 287, 1960–1964

30 Roses, A.D. (2000) Pharmacogenetics and the practice of medicine. *Nature* 405, 857–865

31 Adler, E. *et al.* (2000) A novel family of mammalian taste receptors. *Cell* 100, 693–702

32 Beeley, L.J. *et al.* (2000) The impact of genomics on drug discovery. *Prog. Med. Chem*. 37, 1–43

33 Fukusumi, S. *et al.* (1997) Identification and characterization of a novel human cortistatin-like peptide. *Biochem. Biophys. Res. Commun*. 232, 157–163

34 Bafna, V. *et al.* (2000) Ligand-receptor pairing via tree comparison. *J. Comp. Biol.* 7, 59–70

35 Southan, C. (2000) Assessing the protease and protease inhibitor content of the human genome. *J. Pept. Sci.* 6, 453–458

36 Gill, R.W. *et al.* (1997) A new dynamic tool to perform assembly of expressed sequence tags, ESTs. *Comp. Appl. Biosci*. 13, 453–457

37 Prigent, C. *et al.* (1999) *In silico* cloning of a new protein kinase, Aik2, related to *Drosophila aurora* using the new tool: EST Blast. *In Silico Biology* 1, 123–128

38 Hayes, P. *et al.* (2000) Cloning and functional expression of a human orthologue of rat vanilloid receptor-1. *Pain* 88, 205–215

39 McKusick, V.A. (1998) *Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders*. (12th edn), Johns Hopkins University Press

40 Marcotte, E.M. *et al.* (1999) A combined algorithm for genome-wide prediction of protein function. *Nature* 402, 83–86

41 Roses, A.D. (1994) Apolipoprotein E affects the rate of Alzheimer's disease expression: beta-amyloid burden is a secondary consequence dependent on APOE genotype and duration of the disease. *J. Neuropathol. Exp. Neurol*. 53, 429–437

42 Lai, E. *et al.* (1998) A 4-Mb high-density single nucleotide polymorphism-based map around human ApoE. *Genomics* 54, 31–38

43 Drysdale, C.M. *et al.* (2000) Complex promoter and coding region of β2-adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness. *Proc. Natl. Acad. Sci. U. S. A.* 97, 10483–10488

44 Heid, C.A. *et al.* (1996) Real time quantitative PCR. *Genome Res.* 6, 986–994

45 Gibson, E.M.U. *et al.* (1996) A novel method for real time quantitative RT-PCR. *Genome Res.* 6, 995–1001

46 Lockhart, D.J. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Genet*. 14, 1675–1680

47 Schena, M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470

48 Debouck, C. and Goodfellow, P. (1999) DNA microarrays in drug discovery and development. *Nat. Genet*. 21, 48–50

49 Lockhart, D.J. and Winzeler, E.A. (2000) Genomics, gene expression and DNA arrays. *Nature* 405, 827–836

50 Nuwaysir, E.F. *et al.* (1999) Microarrays and toxicology: the advent of toxicogenomics. *Mol. Carcinog.* 24, 153–159

51 Hacia, J.G. and Collins, F.S. (1999) Mutational analysis using oligonucleotide arrays. *J. Med. Genet.* 36, 730–736

52 Hacia, J.G. (1999) Resequencing and mutational analysis using oligonucleotide microarrays. *Nat. Genet*. 21, 42–47

53 Hacia, J.G. *et al.* (1999) Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat. Genet*. 22, 164–167

54 Ermolaeva, O. *et al.* (1998) Data management and analysis for gene expression arrays. *Nat. Genet*. 20, 19–23

55 Golub, T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537

56 Perou, C.M. *et al.* (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. U. S. A.* 96, 9212–9217

57 Perou, C.M. *et al.* (2000) Molecular portraits of human breast tumours. *Nature* 406, 747–752

58 Butte, A.J. *et al.* (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. U. S. A.* 97, 12182–12186

59 Iyer, V.R. *et al.* (1999) The transcriptional program in the response of human fibroblasts to serum. *Science* 283, 83–87

60 Eisen, M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868

61 Tamayo, P. *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2907–2912

62 Hughes, T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell* 102, 109–126

63 Berge, E.B. *et al.* (2000) Accumulation of dietary cholesterol in sitosterolemia by mutations in adjacent ABC transporters. *Science* 1771–1775

64 Spellman, P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9, 3273–3297

65 Zhang, M.Q. (1999) Promoter analysis of co-regulated genes in the yeast genome. *Comput. Chem*. 23, 233–250

66 Wilkins, M.R. *et al.* (1996) From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology* 14, 61–65

67 Pandey, A. and Mann, M. (2000) Proteomics to study genes and genomes. *Nature* 405, 837–845

68 Jensen, O.N. *et al.* (1997) Automation of matrix assisted laser desorption/ionization mass spectrometry using fuzzy logic feedback control. *Anal. Biochem*. 69, 1706–1714

69 Berndt, P. *et al.* (1999) Reliable automatic protein identification from matrix-assisted laser desorption/ionization mass spectrometric peptide fingerprints. *Electrophoresis* 20, 3521–3526

70 Fields, S. and Song, O. (1989) A novel genetic system to detect protein–protein interactions. *Nature* 340, 245–246

71 MacBeath, G. and Schreiber, S.L. (2000) Printing proteins as microarrays for high-throughput function determination. *Science* 289, 1760–1763

72 Uetz, P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627

73 Marcotte, E.M. *et al.* (2000) Detecting protein function and protein–protein interactions from genome sequences. *Science* 285, 751–753

74 Service, R.F. (1999) Wiggling and undulating out of an X-ray shortage. *Science* 285, 1342–1346

75 Palczewski, K. *et al.* (2000) Crystal structure of Rhodopsin: a G protein-coupled receptor. *Science* 289, 739–745

76 Dolyle, D.A. *et al.* (1998) The structure of the potassium channel: molecular basis of K$^+$ conduction and selectivity. *Science* 280, 69–77

77 Shapiro, L. and Scherer, P.E. (1998) The crystal structure of a complement-1q family protein suggests an evolutionary link to tumor necrosis factor. *Curr. Biol.* 8, 335–338

78 Boggon, T.J. *et al.* (1999) Implication of tubby proteins as transcription factors by structure-based functional analysis. *Science* 286, 2119–2125

79 Cho, Y. *et al.* (1994) Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. *Science* 265, 346–355

80 Williamson, A.R. (2000) Creating a structural genomics consortium. *Nat. Struct. Biol.* 7 (Suppl.), 953